On the contextual appropriateness of performance rules

R. Timmers (2002), On the contextual appropriateness of performance rules. In R. Timmers, *Freedom and constraints in timing and ornamentation: investigations of music performance.* Maastricht: Shaker Publishing, pp. 85-109.

On the contextual appropriateness of performance rules

Previous research has mainly explained the quality of an expressive performance with respect to its relation to the musical structure. The aim of this study was to show the relevance of the performer in setting the rules for expression. The main hypothesis was that the performer initially sets the tone of the performance and the subsequent expressive variations are evaluated with respect to this expressive norm. Two experiments tested this hypothesis. In the first, 36 listeners rated the quality of the performance of the continuation (second half of the musical stimulus) given the performance of the initiation (first half of the musical stimulus). In the second experiment, 17 subjects rated the aesthetic quality of the six performances of the initiation and of the continuation separately. A regression model was proposed that predicts the quality rating of the first experiment on the grounds of the similarity in rubato extent, velocity pattern, average articulation and average asynchrony between the two segments. The results of the experiments and the model confirmed that the appreciation of an expressive interpretation depends on its relation to preceding expressive variations.

An important part of research into the performance of music has been to make explicit the rules underlying the expressive variations in musical performances. A common observation has been that performers tend to deviate from the duration, intonation and articulation as prescribed in the score (see e.g. Gabrielsson, 1974, 1987; Seashore, 1938; Sundberg, Friberg and Frydén, 1991a) and to make additional variations in dynamics, tempo, articulation and pitch that are not notated but make up the microstructure of the music (Repp, 1990, 1992a; Palmer, 1997).

It has been shown that these variations are far from random. Instead, they are systematic to a considerable degree. Similar variations are made in repeated performances, even after years (Ashley and Trilsbeek, submitted; Clynes and Walker, 1982; Timmers, Ashley, Desain and Heijink, 2000). The variations become greater or smaller when the performer is asked to play with exaggerated expression or with no expression, respectively (Kendall and Carterette, 1990). The variations change in a systematic way in accordance with the performer's interpretation of the music (Palmer, 1989, 1992). And last but not least, similar variations can be found in performances of a single piece by different musicians (Repp, 1992a, 1992b).

It has been suggested that the rules underlying these systematic variations have a systematic relation to the musical structure (Clarke, 1985, 1988; Desain and Honing, 1994; Palmer, 1996b; Repp, 1992a; Todd, 1985), to emotion (Gabrielsson and Juslin, 1996), to gesture (Repp, 1993; Todd 1992), to physical movement (Kronman and Sundberg, 1987; Sundberg and Verillo, 1980; Todd, 1995) and to mental processing (Repp, 1998a). In other words, the variations are considered to be expressive or meaningful to the extent that they are connected with some more general conception of music.

The generality of the performance rules described in the literature has been considered fairly large. Some of the rules should hold for all performances within a certain style. For example, Clynes (1983) has defined the "composer's pulse", which applies in the performance of every piece of a certain composer in a certain meter. Another example is the rule system of Sundberg, Friberg and Frydén (1991a), which is used to generate performances of Western tonal music (be it a folk tune or a symphony). After some adaptations of the rules, the system can also be used for atonal music (see e.g. Friberg, Frydén, Bodin and Sundberg, 1994).

Performance rules have often been argued to be of a perceptual or generally cognitive nature and therefore generally applicable. In this respect, one might think of the compensatory lengthening and shortening of notes that are perceived as being relatively short or long due to perceptual grouping of rhythmic figures (Penel and Drake, 1998). One might also think of the communicative function of expressive deviations, such as the slowing down at the end of the phrase that marks the phrase boundary (Palmer, 1992; Clarke and Windsor, 2000).

In this context, the main variation between performers is the extent to which one or the other rule is applied, giving rise to different "rule cocktails", or – in a system with fewer rules – the setting of the main parameters (see e.g. Bresin, 1998; Clarke and Windsor, 2000; Kronman and Sundberg, 1987; Todd, 1992).

In the study presented here, a different, more dynamic (i.e. changing, incremental) application of performance rules is explored. The argument is that some rules are defined *within* a performance. This means that these rules are not predetermined or normative for the interpretation of certain music, but are set by the performer in the act of performing. The performer sets up a pattern of expressive variations, which prompts the listener's expectations. Depending on how a performance begins, the expressive variations are expected to continue in a certain way, thus allowing the performer more freedom to set the idea of the performance without the risk of becoming incomprehensible.

The hypothesis investigated in this study applies both to the behaviour of the listener and to the behaviour of the performer. The listener might have some expectations concerning how a

performance of a certain piece might sound, but during the performance these expectations will be adapted in accordance with idiosyncratic (but systematic) treatment of the musical material by the performer. On the performer's behalf, he or she might set a trend, which then should be continued till a new gesture or trend can be started. In the process of finding a way or creating an alternative way to express the music, the performer might come up with new strategies, which will set the context for the rest of the performance. Having set the context, later variations might deviate from the initialized norm.

This hypothesis of the dynamics of performance rules has occasionally been mentioned in the literature. Clarke (1995) suggested the possibility that a way of performing certain figures – such as the long/short interpretation of equal quarter notes – can set the norm from which later performances might deviate. This was an interpretation of Desain and Honing (1991), who defined expression within a unit as the deviations of its parts with respect to the norm set by the unit itself. Timmers and Desain (2000) have found musicians referring to the process of setting the norm and deviating from it within the performance of a single piece. Repp (1998a) suggested the existence of expectations on performance variations based on previously heard variations, but rejected this hypothesis on the basis of the findings in his own study. He found that the expected timing deviations related to grouping structure were the same irrespective of the context.

This paper focuses on variations that set the norm that are of another kind than Repp's obligatory expectations that relate to the processing of musical structure. The dynamically established constraints explored in the experiments are typical variations for a performer, and consistency constraints that play a role in the well-formedness of an initiated gesture. In other words, it matters less what gesture or variation is initiated as long as it is treated in a consistent way.

This idea of context-dependent performance rules contrasts with the general treatment of performance rules. It also contrasts with the notion of aesthetic judgements being based on a pre-existing concept of how the piece should be performed (Repp, 1997c). This is not to deny the existence of these global rules but to decrease their importance: they are not the sole determinants of the way a piece of music will be performed and perceived.

Two experiments were conducted to show that the appropriateness of performance variations is context dependent. Both were perceptual experiments in which subjects were asked to focus solely on the way the music is performed. The musical material used was always a fragment of the theme from Beethoven's *Paisiello Variations* for piano solo.

In the first experiment, subjects rated the quality of a continuation (second half of a performance) in the context of different initiations (first half of a performance). In the second experiment, subjects rated the quality of the continuations and of the initiations separately. A comparison between the results of the two experiments shows the extent to which the rating of the combined performance (experiment 1) agrees with the rating of the segments of the performance (experiment 2). Little agreement between the experiments and an interaction between the rating of the continuation and its context in experiment 1 would provide evidence in favour of context-dependent performance appreciation. A model is proposed that predicts the quality rating of the continuation of the combined performances based on the similarity between the two segments of the stimuli. A regression analysis provides insight into the performance aspects that played a role in the judgements.

Experiment 1

Method

Subjects

Thirty-six subjects participated in the experiment. Four were music cognition researchers who play an instrument at an advanced level; the others were graduate music students (N = 9) or professional musicians (N = 23). They were selected on the sole ground of being good performers and therefore presumably having good music listening skills and a clear feeling for rules governing a performance of a classical piece. The group comprised pianists (N = 16) and non-pianists (N = 20).

Stimuli

The materials used in the experiments were audio recordings of performances of selected fragments of the theme of Beethoven's *Paisiello Variations* for piano solo (see figure 1). These

recordings were made of performances on a Yamaha Disklavier Pro C3 grand piano collected in a previous study (see chapter 6 (Timmers, Ashley, Desain, Honing, & Windsor, in press))[1].

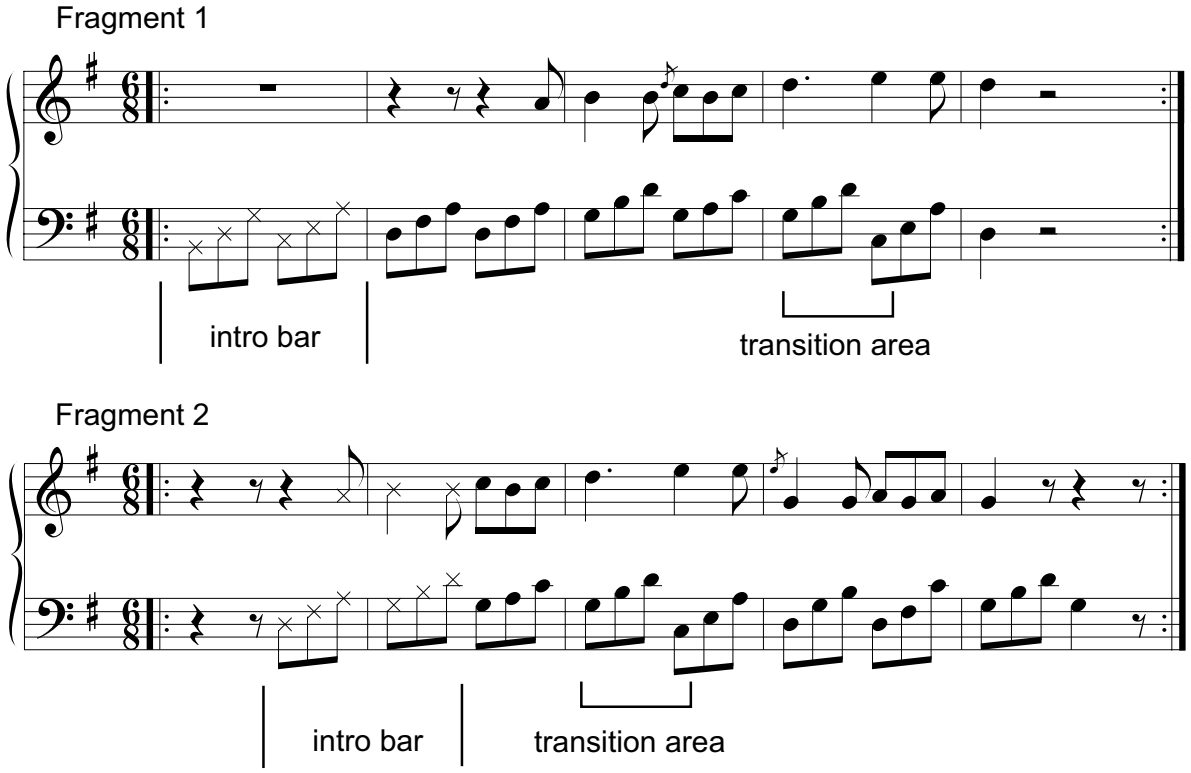

Figure 1. Score of fragments 1 and 2 of the grace note study of chapter 6 (Timmers et al., in press). The intro bar was performed by a computer to indicate the tempo. The subjects performed the rest of the fragment.

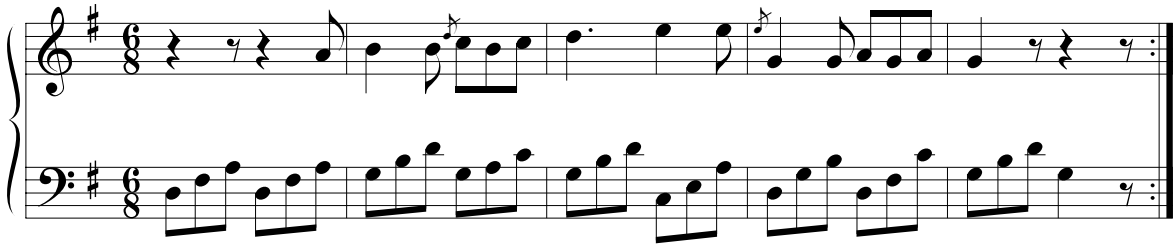

Figure 2. Score of the performances used in experiment 1. It consists of two halves: an initiation (mm. 1-3.5) and a continuation (mm. 3.5-5).

Performances by six professional pianists of fragments 1 and 2 were selected to serve as material for the current study. Performances of measures 2-4.5 of fragment 1 (Figure 1, top) were combined with performances of measures 3.5-5 of fragment 2 (Figure 1, bottom) in all

[1] The performances can be found on www.nici.kun.nl/mmm/ under the header demo's.

possible ways to form performances of the entire phrase (see Figure 2). This resulted in 36 performances of the phrase (six times six first and second halves).

One benefit of using the data collected in the experiment of chapter 6 (Timmers et al., 2002) was that the performances had similar tempi. In the experimental conditions under which the performances were recorded, the pianists were asked to perform in a certain global tempo. The tempo was reinforced by a metronome followed by a computer-generated performance of one bar. The pianists performed the musical fragments immediately after the computer stopped, thereby continuing the performance of the computer. Another benefit was that the dataset consisted of performances of the same fragments by 16 professional pianists. This made it possible to select performances that were especially well suited for the current experiment.

The selection of performances was made on the grounds of having certain striking features concerning rubato, dynamics, articulation, grace note duration and/or asynchrony. The performances had to be characteristic and needed to have something distinctive, although some performances had to be alike as well. The selection led to performances that made good, less good and bad combinations.

Table 1 is an overview of the characteristics of the selected performances. Pianist 1 is mainly on the lower side of the range. Pianists 2 and 4 show clear phrasing and have an intermediate to long grace note. Pianist 3 is generally in between; he has intermediate articulation and asynchrony and small to intermediate rubato. He has some marked contrasts between the two segments, especially for the grace note that is short in the first half and long in the second half. Pianist 5 has minimum values for all dimensions, except for the grace notes, both of which are fairly long. Pianist 6 is mainly on the high side of the range, except for the grace notes, which are relatively short.

The taxonomy given in Table 1 characterizes each performance in a global way. It specifies the average amount of rubato of each half (std. deviation of the accompaniment IOI), the average articulation ratio (the ratio between the duration of the accompaniment notes and the accompaniment IOI) and the average asynchrony between the melody and the accompaniment. These global measures were taken as such, because in the experiment the subjects were supposed to generalize the characteristics from the first half to the second half.

The measures mainly concern the accompaniment, since this is the voice that sounds throughout the fragment and it has the same note duration throughout the fragment. The relation between the melody and accompaniment was characterized by the asynchrony. The grace note is the most influential factor on the asynchrony. Because the grace note is occasionally performed on the beat, some of the values have become negative.

The exceptions to this general way of characterizing the performances were the grace note duration and the velocity pattern. The former is the time interval in milliseconds between the grace note onset and the next melody note onset, and the latter is the characterization of the velocity envelope as being flat, rising or falling. This characterization separates the performances with a clear dynamic phrasing (a rise in dynamics towards the middle of the phrase and a fall in dynamics towards the end of the phrase; see e.g. Todd, 1992) from those without such clear phrasing.

Table 1. *Taxonomy of the 12 performances; amount of rubato measured as the standard deviation of the accompaniment IOI. Velocity pattern (rising, falling or flat). Average articulation measured as the ratio between the duration and the IOI of the accompaniment. Grace note duration is measured as the time interval between the onset of the grace note and the onset of the next melody note. Asynchrony between melody and accompaniment notes in ms. A positive asynchrony means that the melody note is earlier than the accompaniment note, while a negative asynchrony means that the accompaniment leads on the average.*

Pianist	Rubato (std dev)	Velocity (Pattern)	Articulation (Dur/IOI)	Grace IOI (ms)	Asynchrony (ms)
1					
initiation	15.6	Flat	0.978	47	5
continuation	17.3	Flat	1.118	96	3
2					
initiation	18.8	Rise	1.072	116	6
continuation	28.2	Fall	1.076	170	-6
3					
initiation	12.1	Flat	1.163	60	11
continuation	35.5	Flat/fall	1.104	157	-8
4					
initiation	14.2	Rise	0.968	36	16
continuation	35.2	Fall	0.980	114	14
5					
initiation	16.7	Flat	0.916	134	-16
continuation	15.1	Flat	0.959	119	-16
6					
initiation	38.9	Rise	1.268	49	24
continuation	77.3	Fall/flat	1.293	74	27

The differentiation between these envelopes was made on the basis of regression analyses. For each voice of each segment of each pianist, a line was fitted to the velocity of notes with increasing score time. If this line fit showed a significant increase or decrease of velocity with increasing score time, the velocity pattern was assigned a rising or a falling pattern, depending on the direction of the line fit. If there was no significant fit, the pattern was

assigned to be flat. The rising patterns only occurred in the first half, while the second half only showed falling velocity slopes.[2]

Note that a second-order line fit would have yielded the same results this linear line fit provided. The performances without a significant fit showed no trend in the overall velocity level. Their velocity would be better described as a saw-tooth pattern.

Stimuli construction

To combine the performances, the global tempo and average velocity of pairs of performances were adjusted in such a way that the transition between the performances would not give rise to a sudden tempo or velocity change. The overall tempo of fragments 1 and 2 was adjusted such that the duration of the accompaniment of the transition area (see Figure 1) became the same for the first and the second performances. The velocity was similarly adjusted so that the average velocity of the melody and accompaniment notes of the transition area became the same for the performances of the first and the second fragment.

[2] Surprisingly, only the velocity profiles could be easily characterized. The IOI profiles showed in general much less clear patterns.

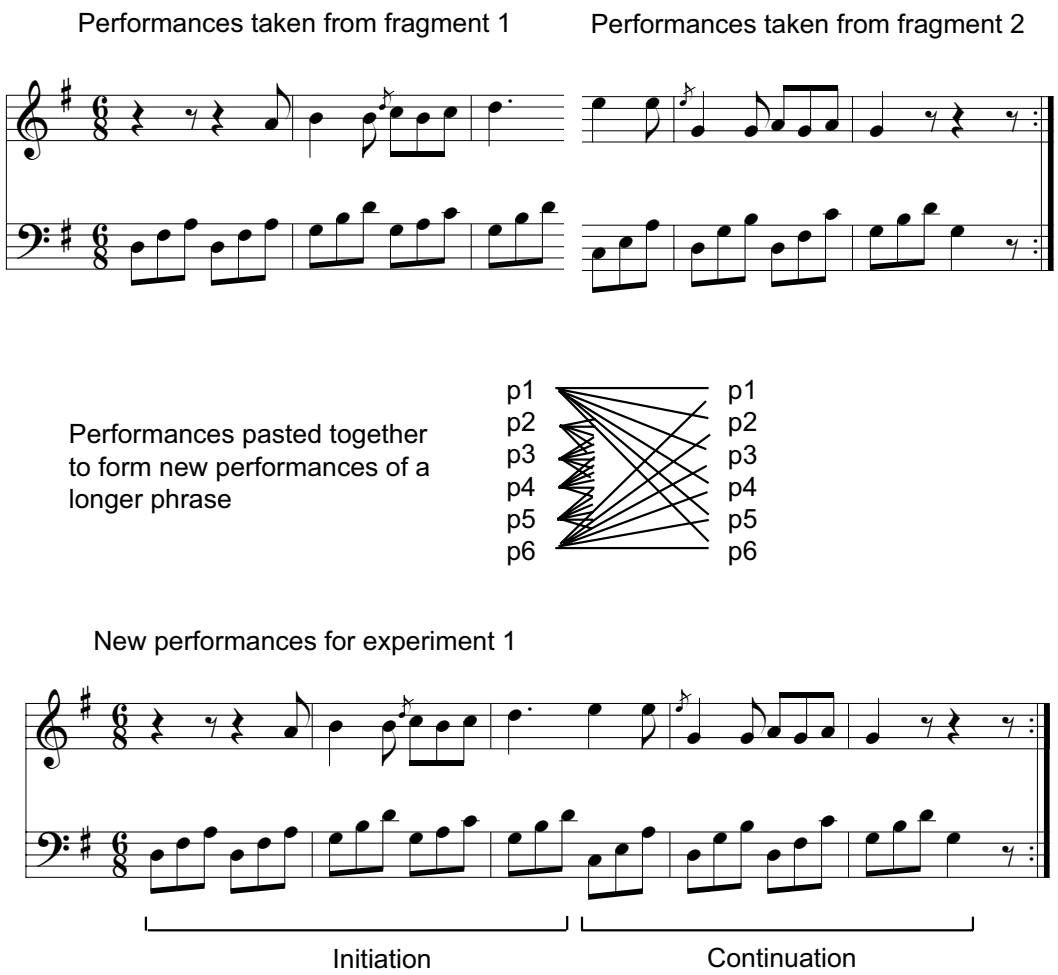

Figure 3. Combination of performances: a performance of fragment 1 is concatenated to a performance of fragment 2. The combinations (p1 with p1, p1 with p2, etc.) make up new performances of a longer fragment; the phrase used in experiment 1.

The adaptations in tempo that had to be made were minor (all below the typical JND of 5%). The adaptations in velocity ranged from a velocity boost of 0 to 7 MIDI velocity units with an average of 2.8. An adaptation of 7 MIDI velocity units is considerable, especially with respect to the average velocity – which was generally around 42 – and the velocity range, which was between 20 and 45 velocity units.

This way of combining performances thus gave rise to initiations and continuations that differed somewhat in average loudness in different conditions; in other words, an initiation or continuation was performed more loudly in some combinations than in others. The effect of these global velocity differences on the results is probably negligible, since the similarity between the performances of a single pianist outweighs the differences caused by velocity level, and the judgement in this experiment consisted of a comparison between the first and second halves.

The MIDI data of the two transformed performances were then concatenated to form performances of an entire phrase (see Figure 2). The MIDI data of measures 1-3.5 were taken from transformed performances of fragment 1 and the MIDI data of measures 3.5-5 were taken from transformed performances of fragment 2. The concatenated MIDI performances were played back on the MIDI grand piano and the sound was recorded using Opcode Vision DSP. The transformation of MIDI data and the construction of stimuli were done using POCO (Honing, 1990) and JMP 3.2.2. Figure 3 explains the combination of performances graphically.

Procedure

The subject was seated in front of a portable Macintosh computer. She/he read the instructions from a text file and heard the sound over headphones. The instructions indicated that the subject would hear performances of a fragment of the theme of Beethoven's *Paisiello Variations* and that she/he should to give an aesthetic judgement of the continuation given the performance of the initiation. For a group of six performances, the first half of the phrase would each time be performed in the same way, while the second half of the phrase would each time be performed differently. The first half was given separately in order to familiarize the subject with the standard.

The user interface showed seven play buttons on a screen. One button contained only the initiation. The six other buttons contained performances of the entire phrase (see Figure 2). These six whole phrases had the same initiation but different continuations, that is, the initiation of the phrases was performed by the same pianist every time, but each continuation was performed by a different pianist.

The subject alternately listened to the first half only and to a complete performance, and then rated the performances on a scale of 1 to 7 (1 = bad continuation, 7 = good continuation). The rating was registered by clicking on the radio button indicating the number (see Figure 4).

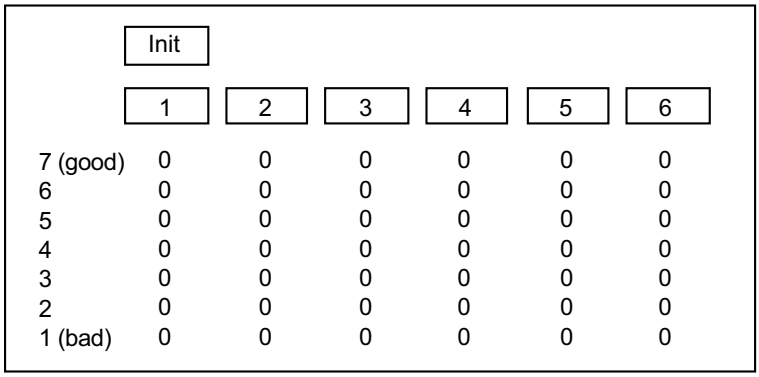

Figure 4. User interface for experiment 1. At the top, the play buttons for the performances (one initiation only and six entire phrases). The rating was registered by clicking a radio button below the performance.

There were a total of six groups of six performances with each group having a differently performed initiation. The order of the initiations was randomized. The order of the continuations that were combined with the initiations was also randomized. The user interface and playback system were made in POCO (Honing, 1990).

Results

Most subjects took considerable time over the first group of six performances, but passed judgement more quickly in the following groups of performances. The total duration of the experiment was 30 to 40 minutes. The subjects reported that the task was difficult at first, then easier as they became more sensitive to the differences, and then difficult again as they became tired.

The subjects also reported that they found it difficult to judge the relation between the first and second half objectively, and instead often made the judgement more subjectively in the sense that they judged the quality of the second half (or the whole) intuitively, without making an explicit comparison.

If the subjects indeed put less effort in comparing the first and second half, this would bias the experiment towards results in which the context is of less influence than the hypothesis implies. If they just focussed on the quality of the second half, this would give rise to a strong effect of continuation, without an interaction with context. If they focussed on the quality of the whole, their aesthetic judgement would be based on the combinatorial quality of the initiation and the continuation, which might be additive or interactive.

In both cases, the bias of the strategy strictly works against the hypothesis of context-dependent appreciation of the continuation. Nevertheless, it is fairly well possible that the intuitive evaluation of the continuation is unconsciously influenced by the context and that the combinatorial quality of the two segments is interactive rather than additive. If the hypothesis also holds under these other listener strategies, this would strengthen the results.

The consistency between subjects was moderate to low with some positive and some negative exceptions. The average correlation between the ratings of the subjects was 0.32, with a maximum of 0.77 and a minimum of -0.42.

A repeated measures ANOVA tested the effect of the initiation, continuation and the interaction between initiation and continuation on the rating of the goodness/fittingness of the continuation (see Table 2). There was a significant effect of continuation and a significant interaction between initiation and continuation.

Table 2. *Results of a repeated measures ANOVA with quality rating of continuation as dependent variable and initiation, continuation, and initiation*continuation as within subject factors.*

Independent variable	Rating of continuation as dependent variable	
Initiation	F (5, 31) = 0.078	p = 0.995
Continuation	F (5, 31) = 33.286	p < 0.001
Initiation * continuation	F (25, 11) = 3.028	p = 0.029

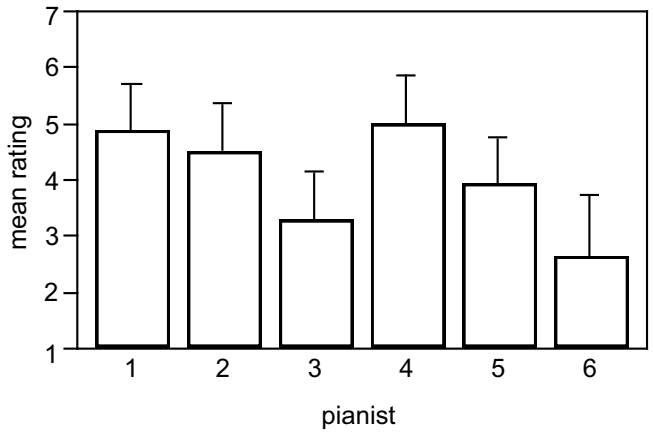

Figure 5. Average quality and standard deviation of continuations 1-6 averaged over context.

The non-significant effect of initiation indicates that the initiations did not bias the rating. Instead, the subjects generally used the entire scale for each initiation (see figure 6 below).

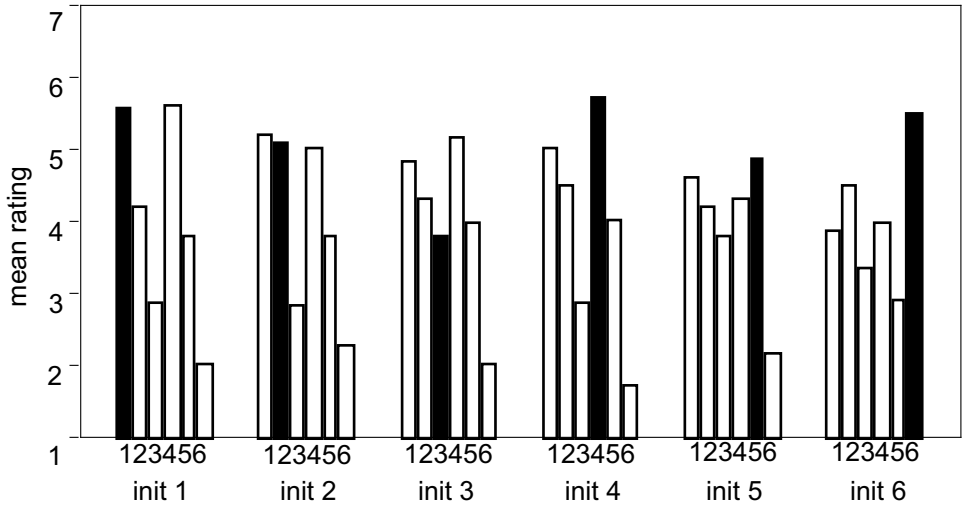

Figure 6. Average quality of continuations 1-6 in context of initiations 1-6.

Figure 5 plots the overall effect of continuation on the rating. It shows that continuations 1 and 4 were rated high, on average, higher than continuations 3 and 6.

Figure 6 shows the effect of the interaction between initiation and continuation on the rating of the continuation. It can be seen that for most initiations, one or two continuations are rated highest. These peaks often include their own continuation (i.e. the combinations marked by a dark colour) with the exception of continuation 3, which was rated low in all conditions. It was rated relatively high in the context of initiation 3.

Most continuations have one condition (or sometimes two conditions) in which they were rated highest. Especially continuation 6 was rated low in all contexts except that provided by initiation 6. Continuations 1 and 4 were rated high in several conditions, which agrees with the higher average of these continuations.

A separate analysis of the effect of initiation on the rating of the continuation provides more insight into the interaction demonstrated above. Therefore, six repeated measures ANOVAs were conducted to test the effect of the initiation given a certain continuation (see Table 3).

Table 3. *Results of six repeated measures ANOVAs that tested the effect of initiation (within subject factor) given a certain continuation.*

Rating cont 1	F (5, 31) = 5.808	p = 0.001
Rating cont 2	F (5, 31) = 1.430	p = 0.241
Rating cont 3	F (5, 31) = 2.315	p = 0.067
Rating cont 4	F (5, 31) = 5.541	p = 0.001
Rating cont 5	F (5, 31) = 5.733	p = 0.001
Rating cont 6	F (5, 31) = 19.429	p < 0.001

For each continuation, with the exception of continuations 2 and 3, there was a significant effect of initiation. A comparison of the six continuations in the context of the different initiations in Figure 6 shows the effects (note that the numbers below the bars indicate the continuations).

Focusing in on continuations 1, 4 and 5 shows that the significant effects of initiation mainly consist of a higher rating of the continuation in their "own" context (in the context of initiation 1, 4 and 5 respectively) and a lower rating in the context of initiation 6. In addition, continuation 4 has a high rating in the context of initiation 1. Continuation 1 has a high rating in the context of both initiation 1 and initiation 2.

Continuation 6 shows a specifically strong effect of context. It was rated high in the context of initiation 6 but low in all other conditions. There was no significant effect of initiation

on continuation 2; it was generally given a rating around 4. There was a slightly higher rating of continuation 2 in the context of initiation 2, but this effect was non-significant.

Nor was there a significant effect of initiation on continuation 3; it generally had a low rating. Only in the context of initiations 3 and 5 was it rated a little higher than in the other contexts. This effect almost reached significance.

A second way to shed light on the interaction is to examine the rating of the continuations given a certain initiation. Six repeated measures ANOVAs tested the effect of continuation on the rating (see Table 4).

Table 4. *Results of six repeated measures ANOVAs that tested the effect of continuation (within subject factor) given a certain initiation.*

Initiation 1	F (5, 31) = 19.039	$p < 0.001$
Initiation 2	F (5, 31) = 37.716	$p < 0.001$
Initiation 3	F (5, 31) = 13.918	$p < 0.001$
Initiation 4	F (5, 31) = 41.155	$p < 0.001$
Initiation 5	F (5, 31) = 11.499	$p < 0.001$
Initiation 6	F (5, 31) = 8.566	$p < 0.001$

The groups of continuations per initiation in Figure 6 show the effects. It is clear that for each initiation there are continuations that were more appreciated and more appropriate than other continuations. Initiations 1, 2 and 4 show an especially neat pattern of well and less well appreciated continuations. For these initiations, continuations 1, (2,) and 4 were rated high, while continuations 3 and 6 were rated low.

The pattern for initiations 3 and 5 is much less pronounced. Most continuations were rated medium high. Only continuation 6 was clearly judged bad in the context of initiations 3 and 5.

Initiation 6 shows a high appreciation of continuation 6 and a medium to bad appreciation of the other continuations. Continuation 2 is relatively strong in the context of initiation 6.

To summarize, for each initiation there were continuations that were considered much better or much worse than others. Often the continuations that were rated high were the same ones in different conditions, as were the continuations that were rated low. Note, however, that there was a significant interaction between the rating of the continuation and the context in which it occurred: continuations 1 and 4 were high for all initiations except initiation 6. Continuation 3 was always rated low. Continuation 6 was appreciated in the context of initiation 6 but not in the other contexts. Continuations 2 and 5, finally, were not markedly preferred or

disliked in most contexts, except that continuation 5 was preferred in the context of initiation 5 and disliked in the context of initiation 6.

Experiment 2

In experiment 1, subjects were asked to rate the quality of the continuation given the initiation. The hypothesis was that the quality of the continuation would depend on its relation with the initiation. And, indeed, the analyses showed an interaction between the rating of the continuation and the initiation preceding it. There was, however, a stronger effect of continuation in the sense that some continuations were on average preferred above other continuations. Further, the subjects reported that sometimes they had difficulties in judging the continuation with respect to the initiation and instead judged the quality of the continuation (or the whole) more intuitively. This might imply that most of the results of experiment 1 were due to context-independent evaluation of the continuation and the initiation, and only a minor part was due to context-dependent evaluation of the continuation.

To gain a better insight into the context dependence or context independence of the judgements, a second experiment was conducted. In it, the subjects were asked to judge the quality of the initiations and continuations separately (i.e. without a context being presented). A comparison between the results of experiment 2 and those of experiment 1 provides an indication of the extent to which the results of experiment 1 were or were not due to general, context-independent appreciation of the performances.

Method

Subjects

Seventeen subjects participated in experiment 2. They were a subset of the subjects of experiment 1. Most were professional musicians (N = 11); the others were graduate students (N = 3) or music cognition researchers (N = 3). There were seven pianists and 10 subjects who played other instruments.

Stimuli

The stimuli of experiment 2 were the six performances of the first half (mm 1-3.5, Figure 2) and the six performances of the second half (mm 3.5-5, Figure 2) used in experiment 1 to form performances of the entire fragment. In this experiment, the two halves were given separately in blocks that first contained all initiations and then all continuations.

To make the performances more easily comparable, the global duration of the first and second halves was normalized to a common value. This was done by adjusting the IOI between notes so that the total duration of the first half was always 4.53 seconds (see Table 5). The duration of all notes was multiplied by the same factor to adjust the articulation to the new tempo. The tempo and duration change always remained below 5% (i.e. below the typical JND for duration).

Table 5. *Original length in seconds of the 1st half (initiation) and 2nd half (continuation) for the six pianists, the factor with which the durations (IOI and articulation) of the performances were multiplied to normalize the tempo to a common value, and the old and new tempi of the first bar in 8th notes per minute. The change in tempo remains below the typical JND of 5%.*

	Length (s)		Dur fact	8th note tempo (BPM) bar 1			
	1st half	2nd half		1st half		2nd half	
Pianist				old	new	old	New
1	4.530	4.183	1.00	60.6	60.6	62.8	62.8
2	4.611	4.341	0.98	58.8	60.0	61.9	63.1
3	4.321	4.094	1.05	64.2	61.1	66.7	63.5
4	4.595	4.275	0.99	60.0	60.9	62.5	63.4
5	4.423	4.177	1.02	62.5	61.3	61.2	60.0
6	4.701	4.949	0.96	55.0	57.3	58.0	60.4

Average length 1st half: 4.53 s.

Procedure

The subjects were seated in front of a Macintosh PowerBook computer. The instructions on the screen clarified that the performances used in the previous experiment had been made by combining six performances from the first half with six performances from the second half. They would now hear the first and second halves separately. They were asked to rate the aesthetic quality of the performances on a scale of 1 to 7 according to their taste. The subjects first rated all first halves and then all second halves. The order in which the first halves and second halves were presented was randomized.

Results

The agreement between subjects was moderate to low with positive and negative exceptions. The average correlation between the ratings of the subjects was 0.28, with a maximum of 0.83 and a minimum of -0.63.

A repeated measures ANOVA tested the effect of pianist, segment and interaction between pianist and segment on the rating of quality of the performances (see Table 6). There was a significant effect of pianist (plotted in Figure 7) and a significant interaction between pianist and segment (plotted in Figure 8).

Table 6. *Results of a repeated measures ANOVA with pianist, segment, and pianist*segment as within subject factors and the aesthetic quality rating as dependent variable.*

Pianist	F (5, 12) = 6.99	p = 0.003
Segment	F (1, 16) = 2.48	p = 0.135
Pianist * segment	F (5, 12) = 4.81	p = 0.012

This average quality rating shows that pianist 6 was especially liked and pianist 5 was particularly disliked. The other pianists were not particularly liked or disliked, with pianist 1 being on the low side of the mean and pianist 3 on the high side of the mean.

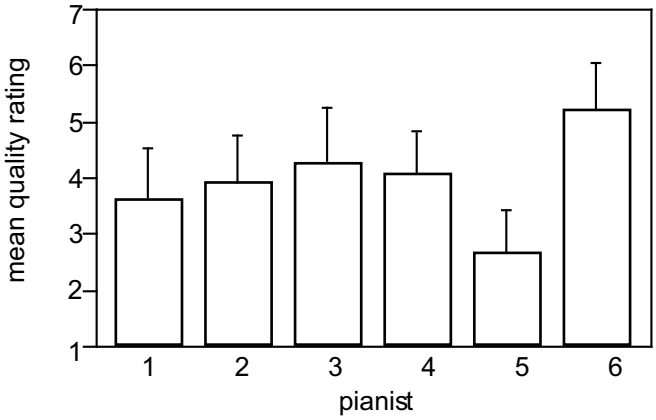

Figure 7. Average aesthetic quality and standard deviation for the six pianists.

Figure 8 shows that there is a marked contrast between the rating of the first and second segment of pianists 3 and 5, but also for pianists 1 and 2. For pianist 3, the initiation was highly appreciated but the continuation was disliked. For pianists 1, 2 and 5 the continuation was liked more than the initiation. Still, the ratings of these second halves were not very high, but liked only moderately. There was no difference between the ratings of the two segments for pianists 4 and 6.

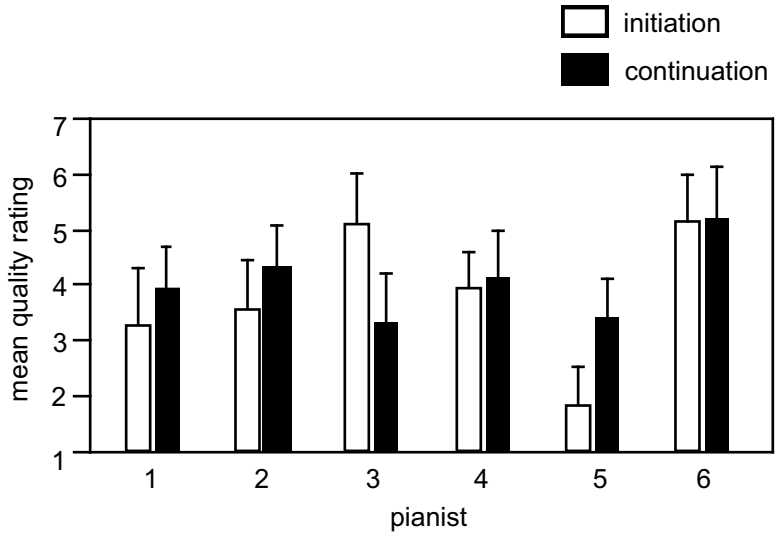

Figure 8. Average aesthetic quality and standard deviation of the two segments of each pianist.

To summarize, initiations 3 and 6 were liked, while initiations 5 and 1 were disliked; only continuation 6 was liked, while continuations 3 and 5 were disliked. The other initiations and continuations were appreciated moderately.

These results do not indicate what exactly was liked or disliked, but only the general preferences. The main significance of these results is with respect to the results of experiment 1 (see discussion).

Comparing results of experiments 1 and 2

Comparing the results of experiment 1 with those of experiment 2 would shed light on the extent to which the results of experiment 1 were due to the subjects' general aesthetic judgements. If the results of experiment 1 were mainly due to the subjects' aesthetic appreciation of the continuation (or the initiation), the ratings of experiment 2 should be fairly good at predicting the rating of experiment 1.

A regression analysis was therefore conducted that took the average rating of the initiations and the continuations of experiment 2 as predictor of the average rating of experiment 1. This model almost reached significance (F (2, 33) = 3.2, $p < 0.06$) and had an R^2 of 0.16. When the model was fitted to the rating of experiment 1 with all between subjects variance taken into account (i.e. not on averaged values), the model did reach significance, but the R^2

dropped to 0.05. Notably, in both regression analyses, only the quality rating of the initiations significantly contributed to the explanation of the rating of experiment 1.

A comparison between the ratings of the continuations given in experiment 1 (plotted in Figure 5) and in experiment 2 (plotted in Figure 8) made clear the discrepancies between the two ratings of the continuation. Most striking is the difference in rating of continuation 6 in the two experiments. While it was given on average a high aesthetic quality in experiment 2, it was given a low rating with respect to most conditions in experiment 1. This clearly means that the low rating of condition 6 in experiment 1 was due to context.

A second clear discrepancy exists between the rating of continuation 1 in the two experiments: it was on average higher in experiment 1 than in experiment 2. This indicates that although initiation 1 was not particularly liked, it was still appropriate in the context of several initiations.

To see whether the model might do better for individual subjects, the same regression analysis was done for the subjects that participated in both experiments. For these 17 subjects, the regression analysis reached significance in nine cases. The explained variance varied between 0% and 44%. The average R^2 was 0.21. Of the nine cases for which the model was significant, the contribution of the initiation was significant five times and the contribution of the continuation was significant six times.

To summarize, in general, the data of experiment 2 are not a strong predictor of the data of experiment 1, which indicates that the appreciation of the combined performances is badly explained by the appreciation of the segments. Part of the effect of continuation in experiment 1 (e.g. the on average high rating of continuation 1 and the on average low rating of continuation 6) was due not to general aesthetic preference but to context.

Only for some subjects did the data of experiment 2 explain a major part of the results of experiment 1. Those subjects seem to have based the rating of experiment 1 to a considerable extent on their appreciation of the initiation and the continuation.

Model

Experiment 1 showed that the appropriateness of expressive variations in the second half of the Beethoven phrase depended on the expressive characteristics in the first half of the phrase. Experiment 2 gave further evidence that for most subjects, context (i.e. initiation) was the important factor for the judgements of experiment 1. The question remained on what grounds a performance was considered to be a good continuation of the initiation, and what aspects of the initiation were expected to continue in the continuation.

Similarity measures

On the grounds of the descriptions of the stimuli given in the method section of experiment 1, it was possible to formulate a model that predicts the consistency in expression of a performance combination. This model assumed that the quality judgement of the combined performances of experiment 1 related directly to the similarity of the two performances along five dimensions: (a) amount of rubato, measured as the standard deviation of accompaniment IOI; (b) velocity pattern, which is either flat or with a crescendo in the first half and a decrescendo in the second half; (c) average articulation, measured as the duration/IOI ratio of the accompaniment; (d) grace duration, measured as the time interval between grace onset and next melody note onset; and (e) asynchrony, measured as the melody note onset time minus the accompaniment note onset time.

The similarity measures were calculated as follows. First, the average value of the parameters was calculated for each half (see Table 1). Then, for each performance combination, the absolute difference between the rubato, articulation, grace note and asynchrony values of the halves was calculated. Finally, these difference values were inverted to similarity measures in a range between 0 and 1, where 0 was the most different combination and 1 indicated an exact similarity between the halves.

Formulas 1-4 show the definitions. A capital *D* means the difference measure of rubato (*rub*), articulation (*art*), grace duration (*grace*) and asynchrony (*asyn*), respectively.

$$D_{rub} = \left| \overline{std(IOI)_1} - \overline{std(IOI)_2} \right| \quad (1)$$

$$D_{art} = \left| \overline{dur/IOI_1} - \overline{dur/IOI_2} \right| \quad (2)$$

$$D_{grace} = \left| graceIOI_1 - graceIOI_2 \right| \quad (3)$$

$$D_{asyn} = \left| \overline{asyn_1} - \overline{asyn_2} \right| \quad (4)$$

Formula 5 shows the calculation of the similarity measure. A capital *S* means the similarity measure, which is derived from the difference measure. The subscript to the measure indicates the dimension of the calculated similarity.

$$S_x = 1 - \frac{D_x}{\max(D_x)} \quad (5)$$

The similarity between velocity patterns was measured in a different way. The combinations of patterns were given a hierarchy. There were four possible combinations: rise-fall, rise-flat, flat-fall and flat-flat. Of these, the combinations rise-fall and flat-flat can be considered consistent interpretations, or combinations with similar velocity treatment, in the sense that rise-fall signals a performance with clear phrasing in both halves and flat-flat signals a performance without clear phrasing in both halves. In this line of reasoning, these two combinations have a high similarity rating, while the other two combinations have a low similarity rating.

These consistency considerations were complemented with a preference for phrasing in velocity above no phrasing. Todd (1992, 1995) and Friberg, Sundberg and Frydén (1994) have argued and demonstrated that a well-shaped phrasing contains an increase in velocity towards the middle of the phrase and a decrease in velocity towards the end of the phrase, paralleling the acceleration and deceleration in tempo. Clarke and Windsor (2000) have further shown that a decrease in velocity especially signals phrase endings.

The ordering of pattern combinations was therefore, from best to worst: rising-falling (1.00), flat-falling (0.67), flat-flat (0.33) and rising-flat (0.00). The intermediate cases in which the fall was only present in one voice were given an intermediate value. The effect of context is the penalty for a flat second half after a rise in velocity in contrast to the higher rating of a flat continuation that follows a flat initiation.

Validation

A regression model was fitted to the judgement data of experiment 1 and optimal parameter settings were calculated for the different factors. A prediction of quality rating was made on the

grounds of these factors and these parameter settings. A comparison between predicted data and averaged data (see Figure 9) shows that the model comes very close to the observed values. The correlation between the predicted values and the observed values is 0.82.

The model is too optimistic for the combinations that include continuations 3, which can be understood with respect to the results of experiment 2 that indicated that continuation 3 was particularly disliked. It is too pessimistic, on the other hand, for certain combinations that include continuations 1 and 4. This might suggest that some differences between the halves in combinations including continuations 1 and 4 are not evaluated that strongly by the subjects. The subjects' positive evaluation of these combinations might also be due to a lack of better options; they are the best combination, given the set of performances, and might therefore be rated highly.

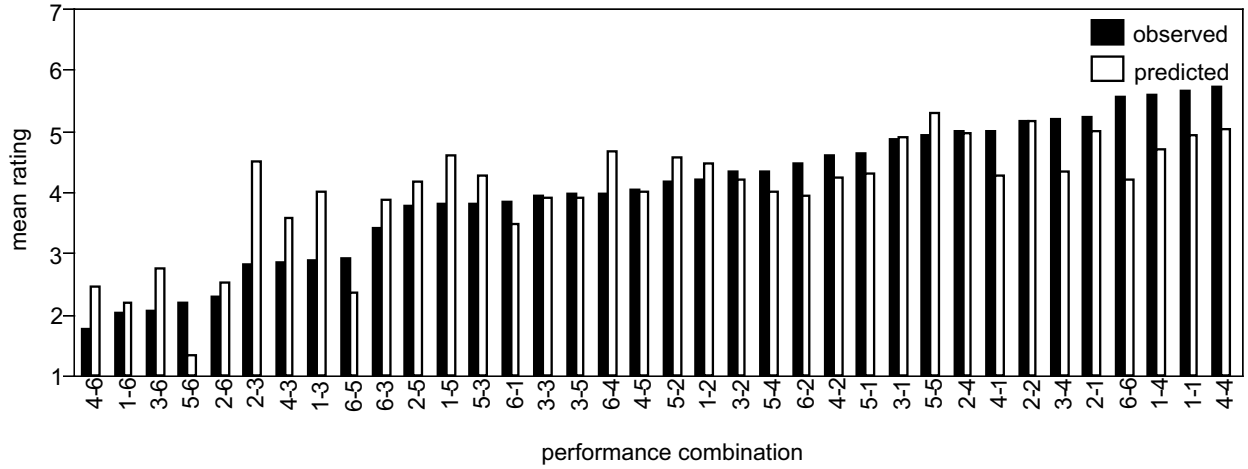

Figure 9. Predicted rating and observed rating of experiment 1 per performance combination.

The multiple regression analysis further tested the contribution of each factor to the explanation of the ratings of experiment 1. All factors except the grace note duration reached significance (see Table 7). The overall R^2 was 0.22 when all between subjects variance was taken into account. This indicates that, although the factors are reasonably strong in explaining the variance of the rating data, the between subjects variance is nevertheless considerable.

The same regression analysis was therefore fitted for each subject separately. Appendix 1 shows the R^2 obtained in this analysis as well as the p values of the whole model and the factors that contributed significantly to the fit.

For most subjects, the model was fairly good at explaining the ratings. For 13 subjects (out of 36 subjects), the explained variance was even over 50%. For five subjects, however, the model did not reach significance. The average R^2 using this method was 0.43, which is

considerably higher than the average R^2 of 0.21 obtained in the regression between individual results of experiments 1 and 2.

Table 7. *Results of a regression analysis with similarity in rubato, velocity pattern, articulation, grace duration and asynchrony as independent variables and rating of experiment 1 as dependent variable.*

Rubato	F (1, 1326) = 108.22	p = 0.000
Velocity pattern	F (1, 1326) = 13.58	p = 0.000
Articulation	F (1, 1326) = 17.67	p = 0.000
Grace duration	F (1, 1326) = 2.29	p = 0.130
Asynchrony	F (1, 1326) = 95.60	p = 0.000

Note that in most cases only one factor contributed significantly to the explanation, which suggests that the subjects generally paid most attention only to one dimension. There were only three subjects for which three factors reached significance.

Rubato was the factor that most often reached significance (for 20 subjects). Asynchrony was significant for nine subjects; grace duration was a factor for six subjects; articulation was a factor for five subjects; and velocity was a factor for our subjects (and almost reached significance for two other subjects (p = 0.06)).

For the subjects that participated in both experiments, the R^2 obtained with this similarity model was always higher than the R^2 obtained with the aesthetic quality model of experiment 2. So even for the subjects for whom the aesthetic quality model worked well, the similarity model worked better.

The average R^2 for the 17 subjects was 0.46 (as opposed to 0.21 for the aesthetic quality model). When the similarity model was extended with the ratings of the initiations and continuations of experiment 2 and fitted to the ratings of experiment 1, the average R^2 rose to 0.54.

To summarize, the extent to which the continuations of experiment 1 were appreciated can be fairly well predicted by a model that takes into account the similarity in rubato extent, velocity pattern, articulation, grace duration and asynchrony between the performances of the two segments. A similar rubato, articulation, grace duration or asynchrony means a similar amount of these parameters in both halves. A similar velocity pattern means that a rise in velocity is followed by a fall in velocity; it also means that a flat velocity followed by a flat velocity is judged to be better than a rise in velocity followed by a flat velocity.

Different subjects focused on different parameters. The parameter most attended to was the similarity in rubato, followed by similarity in asynchrony.

General Discussion

What these results say about context-dependent performance rules is that the appropriateness of an expressive interpretation depends on the variations that preceded it. For example, a closure of the fragment with considerable rubato and ritardando was appreciated, but only if it had been preceded by an appropriate preparation that already contained rubato and considerable expressive shaping.

The results also showed that an initiation constrains the kind of expressive variation that can appropriately follow it. Given the initiation, certain continuations are expected more or found better suited than others. For example, it is considered inappropriate if a fragment starts off staccato and in tempo, but continues legato and with considerable rubato. And an initiation that has a clear increase in dynamics is not expected to be followed by a continuation with flat dynamics; instead, it is expected to be followed by a decrease in dynamics in order to continue and close of the initiated phrasing.

Most clearly the results have demonstrated a notion of consistency. A performance is expected to continue in a consistent way, without breaks or noticeable transitions in asynchrony, articulation or rubato, at least when it concerns a performance of one phrase that does not contain clear breaks itself. This in turn has demonstrated how regularities in performance can be explained in a non-normative way. It has also suggested the notion of balance: the two parts of the performance should be in balance concerning the amount of rubato and the length of the ornaments.

The generality of the results is limited to the extent that they rely on the set of stimuli used in the experiment. Rubato was found to be the main factor responsible for the judgements in experiment 1. This might partly be due to the fact that the differences in rubato were especially large in the stimulus set.

Another caveat should be made as regards interpreting the asynchrony factor. The differences in asynchrony, besides being the largest differences in the stimuli, mainly signalled differences in treating the ornament as an on-beat or pre-beat *Vorschlag*. The disruption in asynchrony and onset timing the on-beat performance of the ornament caused may have been the main reason for similarity in asynchrony being a factor.

Further, the musical material used in the experiment was simple and therefore made it possible for the subjects to easily attend to the performance variations. The ongoing Alberti bass

made it plausible to expect performance variations to continue in a similar way. It also facilitated the detection of differences between the first and the second half.

The results of experiment 2 are interesting with respect to a study done by Bruno Repp (1997c) on the aesthetic quality of average performances. Though experiment 2 provides too little evidence to draw firm conclusions, comparing the two studies indicates some interesting disparities.

Repp (1997c) found evidence that averaged performances are rated relatively high in aesthetic quality. He also found a negative correlation between individuality and aesthetic quality. In the current study, however, the performance the subjects liked most was the one most different from the other performances. It was the performance that was easily separated from the others and was liked even though it had much rubato for the style (which was mentioned by many of the subjects in the interview following the experiment). Performances (such as those by pianists 1 and 4) that were closer to average treatment of the material (i.e. they fitted more of the initiations well) were less liked. In the interview, the subjects mentioned that there was nothing special or remarkable about these performances.

The reason for these disparities is quite unclear. A possible explanation is that they are due to the more professional subject pool of this study. Repp suggested that the negative correlation between individuality and aesthetic quality that he found might be particularly characteristic for students, while professional performers might stress individuality more.

Numerous directions for future research are suggested by this study. First of all, it would be rewarding to investigate more thoroughly the expected continuation of certain expressive variations. A more systematic and independent manipulation of expressive parameters might give rise to more detailed and formal descriptions of context-dependent performance rules. These rules might include the deviation in amount of rubato that is still allowed; it might also include an expected continuation of a timing or articulation pattern, such as currently is shown only for dynamic phrasing.

Secondly, a production experiment could be a valuable counterpart to this study to see in what ways performers adapt the performance of the continuation in order to fit the context of the initiation.

To conclude, the results of this study are too preliminary to make any inferences on formal descriptions of the rules that are context dependent. The results are, however, strong enough to suggest that performers influence the listeners' framework of interpretation and thus have a tool for determining the expressive boundaries.

Appendix 1

subject	R^2	significant parameters	*p* value
1	0.44	art	0.003
2	0.36	asyn	0.014
3	0.41	rub	0.006
4	0.43	rub, vel	0.003
5	0.40	rub, (vel)	0.007
6	0.30	asyn	0.045
7	0.48	rub, asyn	0.001
8	0.64	grace, asyn	0.000
9	0.52	rub, (art)	0.000
10	0.12		0.539
11	0.50	asyn	0.001
12	0.18		0.295
13	0.27		0.084
14	0.40	rub	0.008
15	0.48	rub, art, asyn	0.001
16	0.61	rub	0.000
17	0.39	rub	0.008
18	0.42	rub	0.005
19	0.58	rub, art, grace	0.000
20	0.64	rub	0.000
21	0.53	rub, vel	0.000
22	0.47	rub	0.001
23	0.61	rub, asyn	0.000
24	0.30	grace	0.046
25	0.52	rub, vel	0.000
26	0.50	vel	0.001
27	0.31	asyn	0.041
28	0.55	rub, art	0.000
29	0.24	(vel)	0.129
30	0.56	rub, grace	0.000
31	0.30		0.049
32	0.45	rub, (asyn)	0.002
33	0.53	rub, grace	0.000
34	0.29	(rub)	0.060
35	0.21		0.187
36	0.39	art, grace,asyn	0.008

Results of regression analyses with similarity in rubato, velocity, articulation, grace duration, and asynchrony as independent variables and rating of experiment 1 as dependent variable. The factors between brackets almost reached significance ($p = 0.06$).